**Summary of Books Analysis project:**

There are millions of books published every year. The same is being purchased regularly. With the increased number of writers, the book is being published around the world in different languages. In the final milestone, the data analysis of the books dataset. The primary intention of the project is to combine the 3 datasets from 3 different sources using the common columns and perform the necessary analysis. Some of the questions as part of my analysis are listed below.

·         The comparison of ratings from different sources.

·         Top 10 books which have been sold and sales in dollars.

·         Which language has the most percentage of books?

·         Under which rating do most books fall.

·         With the increased number of pages, how likely it to get more ratings.

**Here is the source of the datasets:**

Source #1: CSV file Kaggle site.

Source #2: Website data <https://en.wikipedia.org/wiki/List_of_best-selling_books>

Source #3: Api data- <https://www.googleapis.com/books/v1/volumes>

Steps to clean up the data:

CSV file:

The file was read and cleaned.

·         Some of the changes are to change the author’s names from multiple names to a single name.

·         The column names were replaced as required.

·         Some of the columns were not required. So, those were removed.

·         Identify any duplicate records and then remove them.

Website:

Using the Wiki link, the data was retrieved and cleaned.

·         Read the data from the webpage into a dataframe.

·         In the dataframe pick the necessary columns.

·         Replace the column names as required.

·         Drop all the NA records.

·         The columns have repetitive data. So, we removed the repetitive data to have a number alone. Example: 120 million[111] is replaced to 120 and the heading is saved as sold in millions.

API:

Using the good API, the data was retrieved and cleaned.

·         The column contains dates that had values in the format of yyyy or yyyy-mm-dd. To make it consistent we picked the year alone.

·         Drop the NA from certain columns where the data is required.

As part of Milestone 5, I have to combine the data from all the above sources. Then load to the database. Overall, during the course of the projects, I have learned about data cleaning, visualization techniques and SQLite database.